

بخش کتبی

مبحث اول

سوال اول

نادقیقی، تمیز نبودن، نقص داشتن و وجود داده های خارج از محدوده و نا معقول من جمله مشکلاتی هستند که در داده هایی که از دنیا واقعیت به دست می آوریم وجود دارد. به نظر شما برای حل مشکل زیر چه راهکارهایی ارائه می شود؟

- 1) وجود نداشتن یک یا چند ویژگی در داده های آموزش
- 2) نا متعادل بودن توزیع داده ها در کلاس ها
- 3) وجود نویز در داده ها
- 4) وجود ویژگی های correlated (ویژگی هایی که ارتباط زیاد با یکدیگر دارند)

سوال دوم

یک مشاور تحصیلی در حال بررسی روی یک مجموعه داده درباره ساعت مطالعاتی دانشجویان و نمرات آزمون هایشان است. او توانسته است معادله رگرسیون خطی زیر را با توجه به داده های موجود به دست آورد:

نمره آزمون = $60 + 5 * \text{ساعت مطالعاتی}$

اما باتوجه به تاثیر انکار ناپذیر آزمون دادن در آمادگی دانشجویان او قصد دارد که نقش این مسئله را هم در نمره آزمون در نظر بگیرد. به نظر شما او چه مدل ریاضیاتی برای درک ارتباط بین این دو ویژگی پیشنهاد خواهد داد؟ با استفاده از least square method سعی کنید توضیح دهید چگونه ضرایب مناسب را پیدا می کند؟ اگر از gradient descent استفاده کند چطور؟ آیا تکنیک دیگری برای کم کردن اختلاف مجموع مربعات و مقادیر مشاهده شده می شناسید؟

سوال سوم

ارزیابی مدلی که برای پیش بینی استفاده کرده ایم بسیار ضروری است . فرض کنید برای پیش بینی spam بودن از مدل بر مبنای logistic regression زیر استفاده کرده ایم :

$$p(\text{Spam}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * \text{email Length})}}$$

و فرض کنید نتایج به صورت :

TP : 300 , TN :200 , FP :30 , FN : 20

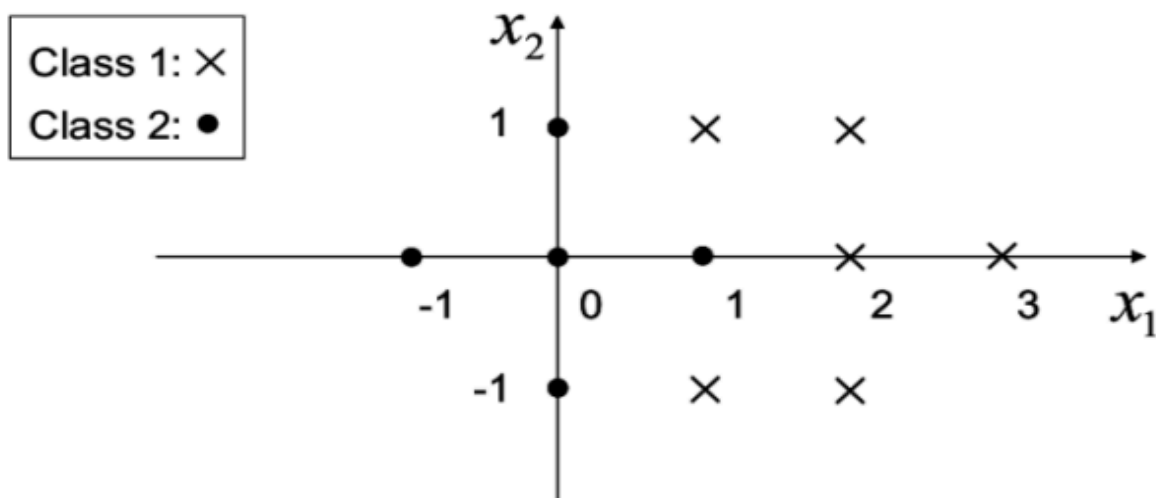
باشد .

برای ارزیابی این مدل ابتدا confusion matrix را رسم کنید ، سپس ، accuracy ، precision ، recall ، F1-score را به دست آورید . توضیح دهید از roc curve چگونه برای ارزیابی این مدل می توان استفاده کرد .

KNN

سوال اول

در تصویر زیر تعدادی نمونه از دو کلاس مختلف مشخص شده اند. داده ی تست (0.5,0) را با روش KNN طبقه بندی کنید .



K نزدیک ترین همسایه با K=3 با دو فاصله ی زیر:

1. فاصله اقلیدسی

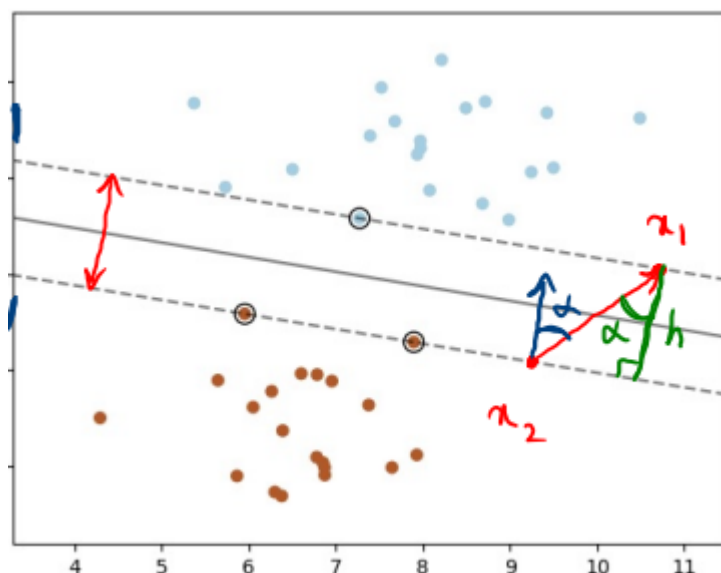
2. فاصله منهتن

Support Vector Machine

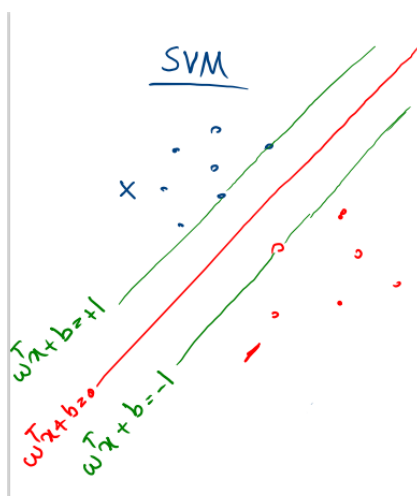
سوال دوم

در این سوال می خواهیم پله به پله به مساله بهینه سازی SVM برسیم و ببینیم معادلات بهینه سازی مقید این طبقه بند از کجا آمده است. همانطور که میدانید در واقع ما در طبقه بند SVM مساله بهینه سازی مقیدی را حل می کنیم که مفهوم پشت آن بدست آوردن خطی است که بیشترین فاصله را از دو کلاس داشته باشد یا به اصطلاح می خواهیم حاشیه (margin) بین 2 کلاس را بیشینه کنیم. برای رسیدن به مساله بهینه سازی SVM مراحل زیر را طی کنید.

1. با توجه به شکل زیر h یا همان margin بین داده دو کلاس را بر حسب W که همان شیب دو خط نقطه چین (فرض کنید خط های نقطه چین شیبی برابر با W دارند.) می باشد بدست بیارید. (راهنمایی: با استفاده از مفهوم ضرب داخلی و با توجه به شکل زیر h را بدست بیاورید.)



2. در واقع تا همین جا کار ما می خواهیم margin که در بخش الف بر حسب W یا همان شیب دو خط می باشد را بیشینه کنیم. پس کافی است شیبی را پیدا کنیم که این حاشیه را بیشینه می کند. اما مساله ما همین جا تمام نمی شود. چرا که این مساله بهینه سازی حاوی قید هایی که باید آن قید ها در نظر گرفته شود. با توجه به شکل زیر سعی کنید قید های مساله بهینه سازی SVM را بنویسید. (راهنمایی: داده های کلاس آبی باید بالای خط سبز بالایی قرار بگیرند و داده های کلاس قرمز باید پایین خط سبز پایینی قرار بگیرند.)



3. در نهایت مساله بهینه سازی SVM که در قسمت 1 بدست آوردید را به همراه قیدهای آن که در قسمت 2 بدست آوردید را کنار یکدیگر بنویسید و درباره روش های حل مساله بهینه سازی مثل روش لاگرانژ تحقیق کنید.(نیاز به حل مساله بهینه سازی نمی باشد).

بخش عملی

مقدمه

هدف این تمرین، آشنایی با روش های یادگیری ماشین¹ جهت پیش بینی قیمت مسکن در شهر بوستون آمریکا است. این تمرین از سه بخش اصلی و یک بخش اختیاری تشکیل شده است؛ در بخش اول به آشنایی با داده ها پرداخته و با توزیع، انواع داده های موجود در مجموعه داده ها و اطلاعات آماری مربوط به مجموعه داده ها آشنا می شویم. به طور کلی این بخش را تحلیل دادگان² می گویند، برای آشنایی بیشتر با چیستی و نحوه عملکرد این بخش می توانید از [این](#) لینک استفاده کنید. بخش دوم که مهمترین بخش در یک پروژه یادگیری ماشینی است، بخش پیش پردازش دادگان است. در این بخش، با استفاده از نتایج و تحلیل های بخش قبلی، دادگان دنیای واقعی را به داده ای قابل پردازش و مناسب برای عملکرد یک مدل یادگیری ماشین تبدیل می کنیم. برای آشنایی بیشتر با این بخش می توانید از [این](#) لینک استفاده کنید.

در بخش سوم به ایجاد مدل های مختلف یادگیری ماشین و در نهایت ارزیابی آن ها می پردازیم. در این بخش که شامل 6 فاز مستقل است. ابتدا به ساخت یک مدل Linear Regression مرتبه اول به صورت

¹ Machine Learning

² Data analytics

دستی(بدون استفاده از مدل آماده) می‌پردازیم، سپس متد گرادیان کاهشی³ و polynomial regression را پیاده سازی کرده و سپس با کمک کتابخانه Scikit-Learn اقدام به تخمین قیمت خانه‌ها می‌کنیم. در فازهای بعدی نیز با مدل‌های پیشرفته‌تر آشنا شده و با استفاده از کتابخانه‌ها به پیاده سازی آن‌ها می‌پردازیم. در نهایت نیز با ارزیابی تمام مدل‌ها استنتاج نهایی را انجام می‌دهیم.

در فاز اول و دوم این بخش لازم است که فایل نوت‌بوک قرار داده شده در سایت را دانلود کرده و بخش‌های مشخص شده را کامل نمایید. پیاده‌سازی فاز سوم نیز در ادامه آن‌ها و در همان نوت‌بوک انجام می‌شود.

لازم به ذکر است تمیزی پیاده‌سازی کدها، استفاده از شیء‌گرایی در پیاده‌سازی مدل‌ها و توابع و دسته بندی منظم آن‌ها حائز نمره امتیازی خواهد بود. برای آشنایی بیشتر با شیء‌گرایی در پروژه‌های یادگیری ماشین می‌توانید از [این](#) لینک استفاده کنید.

بخش اول: آشنایی با مجموعه داده

مجموعه داده‌ای که در اختیار شما قرار دارد، شامل اطلاعات مربوط به قیمت خانه های شهر بوستون به همراه ویژگی‌های خانه‌ها می‌باشد. در این تمرین می‌خواهیم با تحلیل ویژگی‌های خانه‌ها، معیارها و وابستگی‌ها را درک کنیم و بتوانیم قیمت خانه ها را بر اساس ویژگی هایشان درک کنیم.

توضیح ستون های این مجموعه داده در جدول زیر قرار داده شده است:

نام ستون	توضیحات
CRIM	per capita crime rate by town.
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of nonretail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million).
RM	average number of rooms per dwelling.
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers.
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per \$10,000
PTRATIO	pupil-teacher ratio by town
B	$.B_k - 0.63)^2$ where B_k is the proportion of blacks by town)1000

³ Gradient Descent

LSTAT	% lower status of the population
MEDV (Target)	Median value of owner-occupied homes in \$1000s

بررسی مجموعه داده

وقتی یک پروژه یادگیری را شروع می‌کنیم، داده‌هایی که در ابتدا با آن‌ها شروع می‌کنیم، داده‌های خام هستند لذا نیاز داریم که آن‌ها را تجزیه و تحلیل کنیم و یک دید کلی نسبت به داده‌ها به دست آوریم و با ویژگی‌های آن‌ها آشنا شویم. به فاز اولیه تجزیه و تحلیل داده‌ها اصطلاحاً EDA می‌گویند. برای اجرا این فاز گام‌های زیر را انجام دهید:

۱. ساختار کلی داده‌ها را به دست آورید. (برای این کار می‌توانید از متدهای info و describe استفاده کنید).
۲. ممکن است برخی ستون‌های جدول دارای داده‌های از دست رفته باشند، تعداد و نسبت این داده‌ها را به دست بیاورید.
۳. نمودار تعداد مقادیرهای منحصر به فرد برای هر ویژگی را رسم کنید و درباره آن‌ها توضیح دهید.
۴. نمودار وابستگی^۴ ویژگی‌ها به یکدیگر را رسم کنید. درباره این نمودار توضیح دهید و بگویید کدام ویژگی‌ها وابستگی بیشتری به ستون هدف دارند؟
۵. نمودارهای scatter و hexbin معمولاً برای بررسی ارتباط ویژگی‌ها استفاده می‌شوند از این نمودارها برای بررسی وابستگی‌ها با ستون هدف استفاده کنید. کاربرد و چپستی هر یک را مختصراً توضیح دهید.
۶. درباره بررسی‌های دیگری که می‌توانید برای مجموعه داده از آن‌ها استفاده کنید تحقیق کنید و یکی از آن‌ها را پیاده‌سازی کنید.

بخش دوم: پیش پردازش مجموعه داده

مهم‌ترین فاز هر پروژه یادگیری ماشین، فاز پیش پردازش است. در این فاز فرمت داده‌ها را تغییر می‌دهیم، آن‌ها را اصلاح و خلاصه می‌کنیم، تا بتوانیم برای آموزش یک مدل یادگیری ماشین از آن استفاده کنیم. چرا که در دنیای واقعی، اطلاعات جمع‌آوری شده به راحتی کنترل نمی‌شوند و در نتیجه مقادیر خارج از محدوده، ناممکن، از دست رفته و به طور کلی گمراه‌کننده برای آموزش مدل در مجموعه داده‌ها وجود دارند. این فاز باعث می‌شود مدل کارا تری بتوانیم داشته باشیم و سرعت یادگیری بالاتر می‌رود.

۷. روش‌های پر کردن Missing Value را توضیح دهید و حداقل سه روش را پیاده‌سازی کنید. دلیل استفاده از هر روش را مختصراً ذکر کنید.
۸. آیا امکان حذف برخی ستون‌ها وجود دارد؟ چرا؟ در صورتی که این امکان وجود دارد با ذکر دلیل ستون‌های لازم را حذف کنید.
۹. کدام ویژگی‌ها را عددی و کدام‌ها را دسته‌ای می‌گویند؟ تفاوت این دو نوع از ویژگی‌ها در چیست؟ ویژگی‌های عددی و دسته‌ای را در این مجموعه داده‌ها مشخص کنید.
۱۰. در ویژگی‌های عددی، normalizing یا standardizing به چه منظور انجام می‌شود؟ تفاوت این دو روش در چیست؟ در این پروژه نیاز به انجام این کار هست؟

^۴ Correlation

۱۱. برای ویژگی‌های دسته‌ای، که معمولا بصورت یک string یا object در مجموعه داده ذخیره شده‌اند، در آموزش مدل چه پیش‌پردازش‌هایی مفید هستند؟
۱۲. درباره داده های train , validation , test تحقیق کنید و روش های معمول تقسیم بندی را توضیح دهید. سپس دادگان خود را به این دسته‌ها تقسیم‌بندی کنید.
۱۳. درباره سایر روش‌های پیش‌پردازش تحقیق کنید. برخی از این روش‌ها را ذکر کرده و در صورت نیاز از آن استفاده کنید.

بخش سوم: آموزش، ارزیابی و تنظیم

فاز اول: Linear Regression

در این بخش بدون استفاده از کتابخانه سعی کنید که روابط ارائه شده برای رگرسیون خطی را درک کنید و بدون بهره گیری از مدل آماده آن را پیاده سازی کنید.

۱۴. در این پروژه ما در حال پیاده‌سازی مدل‌های یادگیری ماشینی با نظارت هستیم. تفاوت این مدل‌ها با مدل‌های بدون ناظر، نیمه نظارتی و یادگیری تقویتی در چیست؟ برای هر یک مثال بزنید.
۱۵. رگرسیون چیست و چه تفاوت‌هایی با روش‌های دسته‌بندی می‌کند؟
۱۶. روابط ارائه شده در خصوص روش رگرسیون خطی را مختصرا توضیح دهید.
۱۷. بخش های مشخص شده در notebook را تکمیل کنید. از آنجایی که تابع رگرسیون ساخته شده از مرتبه ۱ است، تنها یک ویژگی را می‌توان به عنوان ورودی این تابع انتخاب نمود. به نظر شما کدام ویژگی نسبت به سایر ویژگی‌ها خروجی دقیق‌تری به ما می‌دهد؟ علت انتخاب خود را توضیح دهید.
۱۸. پس از انتخاب ویژگی مناسب از داده‌های train و پیش‌بینی داده‌های آزمون، می‌بایست معیاری برای ارزیابی کارایی خروجی بدست آمده تعیین کنیم. از آنجایی که مدل ما اصطلاحا در حال انجام task رگرسیون است و دسته بندی روی آن انجام نداده‌ایم، نمی‌توان از متدهای ارزیابی کارایی مربوط به classification استفاده کرد. درباره متدهای RMSE, MSE, RSS و R2 score مطالعه کنید و هرکدام را در گزارش خود توضیح دهید.
۱۹. با استفاده از متد RMSE و R2 score، مقادیر پیش‌بینی شده را ارزیابی کنید. عملیات فوق را بر روی چند ویژگی دیگر نیز انجام دهید. از مقادیر بدست آمده چه استنباطی می‌کنید؟
۲۰. مقادیر پیش‌بینی‌شده را با مقادیر واقعی با استفاده از scatter plot مقایسه کنید، که در آن محور x مقادیر واقعی را نشان می‌دهد و محور y مقادیر پیش‌بینی‌شده را نشان می‌دهد. همچنین خط $x = y$ را نیز رسم کنید.

فاز دوم: polynomial Regression

در این قسمت، رگرسیون را با درجات بالاتر انجام می دهیم. در مرحله قبل، توانستیم با استفاده از دو معادله و دو مجهول به مقادیر بهینه وزن ها برسیم. با افزایش درجه به وضوح به لحاظ حل ریاضیاتی دشواری بیشتری متحمل می شویم و همین مسئله اهمیت حل گام به گام و نزدیک شدن تدریجی به وزن های بهینه را نشان می دهد. در نوت بوکی که در اختیارتان قرار گرفته است، یک رابطه برای محاسبه آن با استفاده از ماتریس ویژگی ها بیان شده است.

21. ابتدا با استفاده از رابطه داده شده یک مدل رگرسیون چند جمله ای ایجاد کنید. سپس با استفاده از روش گرادین کاهشی و توابع پیاده سازی شده در نوت بوک یک مدل چند جمله ای بسازید. دقت این دو مدل را با استفاده از متدهای پیشین سنجیده و با مدل رگرسیون خطی مقایسه نمایید.

22. مقادیر پیش بینی شده را با مقادیر واقعی با استفاده از scatter plot مقایسه کنید، که در آن محور x مقادیر واقعی را نشان می دهد و محور y مقادیر پیش بینی شده را نشان می دهد. همچنین خط $x = y$ را نیز رسم کنید.

فاز سوم: طبقه بندی

درخت تصمیم یک مدل پیش بینی است که از ساختار درختی برای تصمیم گیری در مورد مقدار یک متغیر هدف استفاده می کند. این درخت از گره ها و لیستی از تقسیم ها تشکیل شده است که به ازای هر گره، یک متغیر و یک مقدار تقسیم بندی انتخاب می شود تا داده ها به گره های فرزند تقسیم شوند. این فرآیند ادامه پیدا می کند تا ویژگی های مهم مجموعه داده درخت تصمیم را تشکیل دهند. هدف نهایی این است که با استفاده از این درخت، می توان پیش بینی هایی در مورد داده های جدید انجام داد. درخت تصمیم به دلیل قابل فهم بودن ساختار و نتایج آن، یکی از محبوب ترین روش های یادگیری ماشین است.

الگوریتم KNN یا همسایگان نزدیک ترین، یکی از ساده ترین الگوریتم های یادگیری ماشین است که برای دسته بندی و رگرسیون استفاده می شود. در این الگوریتم، تصمیم گیری بر اساس اکثریت آرای همسایه های نزدیک ترین به نمونه ای که می خواهیم طبقه بندی یا پیش بینی کنیم، انجام می گیرد. به بیان ساده تر، KNN با محاسبه فاصله بین نمونه جدید و تمام نمونه های موجود در داده های آموزش، نمونه هایی را که نزدیک ترین هستند شناسایی می کند و بر اساس بیشترین برچسب حاضر در همسایه های نزدیک، برچسب نمونه جدید را تعیین می کند. این الگوریتم نیاز به تنظیم پارامتر K دارد که تعداد همسایگانی را مشخص می کند که در تصمیم گیری شرکت می کنند.

23. مفهوم prune کردن در درخت های تصمیم گیری چیست؟ مزایا و معایب استفاده از این روش را ذکر کنید.

24. استفاده از درخت های تصمیم گیری چه زمانی می تواند نسبت به سایر مدل ها دارای مزیت باشد؟

25. تفاوت ذاتی طبقه بند KNN با سایر روش های طبقه بندی مثل شبکه های عصبی یا Logistic regression در چیست؟ (به نحوه train شدن هر کدام از طبقه بندها دقت کنید.)

26. در رابطه با الگوریتم one nearest neighbor تحقیق کنید و مزایا و معایب آن را ذکر کنید.

27. در رابطه با دیگر روش های سنجیدن فاصله در الگوریتم KNN تحقیق کنید و چند مورد از آن ها را بیان کنید.

28. در این فاز از پروژه، ابتدا همان ستون هدف که شامل میانگین قیمت خانه های تحت اشغال (MEDV) است را به 3 دسته تقسیم می کنیم، دو دهک بالای قیمت ها را خانه های لوکس، دو دهک پایین را خانه های

اقتصادی و باقی خانه‌ها را بعنوان خانه‌های معمولی برچسب‌گذاری می‌کنیم و بعنوان یک ستون جدید ذخیره می‌کنیم. سپس دو مدل بر پایه Decision Trees، K-Nearest-Neighbours با استفاده از کتابخانه scikit learn پیاده‌سازی می‌کنید. سپس فرآپارامترها⁵ را تغییر دهید و مدل را تا حد امکان بهینه کنید. بهینه‌سازی مدل‌ها به این منظور است که تابع هزینه کمینه شود اما overfitting رخ ندهد. یکی از مدل‌های بهینه شده را که با آزمون و خطا به آن رسیده‌اید در گزارش خود نشان دهید.

29. برای هر دو این مدل‌ها با کمک تابع [GridSearchCV](#)، مقادیر بهینه برای پارامترها را بدست آورید. نحوه عملکرد این تابع را به طور مختصر توضیح دهید. و نتایج بدست آمده را با نتایج بدست آمده از مدل‌هایی که فرآپارامترهای آن با آزمون و خطا بدست آمده بود، مقایسه کنید.

30. درخت تصمیم نهایی خود را رسم کنید. (برای این کار می‌توانید از [plot_tree](#) استفاده کنید).

31. آیا در مدل‌های شما underfitting یا overfitting رخ داده است؟ به طور کلی چه زمانی این پدیده رخ می‌دهد؟ هر یک را توضیح دهید.

فاز چهارم: روش‌های Ensemble

روش‌های Ensemble در یادگیری ماشین به مجموعه‌ای از مدل‌ها اشاره دارند که به صورت همکاری برای بهبود دقت پیش‌بینی‌ها کار می‌کنند. این روش‌ها معمولاً با ترکیب چندین مدل ساده‌تر، مدل نهایی را می‌سازند که در مجموع از هر یک از مدل‌های تکی بهتر عمل می‌کند. دو روش اصلی در متدهای Ensemble وجود دارد: Bagging و Boosting به منظور کاهش واریانس مدل‌ها استفاده می‌شود و در آن چندین نمونه از داده‌ها به طور تصادفی انتخاب شده و برای هر نمونه یک مدل ساخته می‌شود. این مدل‌ها سپس ترکیب می‌شوند تا نتیجه نهایی حاصل شود. حال ابتدا برای بررسی بهتر این مفهوم به سوالات زیر پاسخ کوتاه دهید.

32. درباره چرایی استفاده از روش‌های Ensemble و اینکه چرا امروزه این روش‌ها از اهمیت بالایی برخوردار هستند توضیح دهید.

33. مکانیزم کلی روش‌های Boosting و Bagging برای طبقه بندی را تشریح کنید و تفاوت‌های آن را بیان کنید.

جنگل تصادفی یکی دیگر از روش‌های یادگیری جمعی است که بر اساس ایده ای از تجمع از قوانین یا الگوریتم‌های ساده تر، به صورت تصادفی، تعدادی از مدل‌های یادگیری خود را اجرا می‌کند و سپس از ترکیب نتایج حاصل از این مدل‌ها برای پیش‌بینی مقادیر جدید استفاده می‌کند.

در واقع، جنگل تصادفی یک مجموعه از درخت‌های تصمیم است که هر کدام به صورت مستقل از دیگری آموزش داده می‌شوند و سپس نتایج آن‌ها ترکیب می‌شوند تا یک پیش‌بینی نهایی برای داده‌های ورودی انجام شود. این روش برای حل مسائل پیچیده و تعداد زیادی داده بسیار موثر و کارآمد است.

34. نحوه عملکرد روش جنگل‌های تصادفی را به طور مختصر توضیح دهید.

35. مفهوم Bootstrapping در جنگل‌های تصادفی چیست؟ کارکرد آن چگونه است و چگونه بر نتایج مدل تاثیرگذار است؟

36. آیا تعداد درخت‌های تصمیم‌گیری در جنگل تصادفی بر کارایی مدل تاثیر گذار است؟ بهترین مقدار آن به طور تجربی در حدود چه مقادیری است؟

37. استفاده از جنگل تصادفی چه زمانی مناسب نیست؟ این روش در چه زمانی توصیه می‌شود؟

38. استفاده از جنگل تصادفی چه تاثیری روی واریانس دارد؟

⁵ Hyperparameters

39. در این بخش پس از توضیح مختصری درباره هر یک از فرآپارامترهای جنگل تصادفی مجدداً با استفاده از تابع [GridSearchCV](#)، این مدل را آموزش داده و بهترین فرآپارامترها را گزارش کنید. (نیازی به آموزش مدل و انتخاب فرآپارامترها با آزمون و خطا نیست).

امتیازی: XGBoost

در این تمرین قصد داریم تا با الگوریتم XGBoost آشنا شویم. XGBoost یک الگوریتم یادگیری ماشین است که بر پایه روش های گرادیان کاهشی است. این الگوریتم برای حل مسائل مختلف یادگیری ماشین از جمله طبقه بندی، پیش بینی و رتبه بندی مورد استفاده قرار می گیرد. XGBoost قابلیت اجرای سریع، کارایی بالا و افزایش دقت در پیش بینی ها را دارا می باشد.

40. نحوه عملکرد XGboost را مختصراً توضیح دهید.

41. ابتدا مفهوم Gradient Boosting را توضیح دهید و سپس تفاوت بین Boosting Tree و Decision Tree را بیان کنید.

42. در این بخش پس از توضیح مختصری درباره هر یک از فرآپارامترهای XGBoost با استفاده از تابع [GridSearchCV](#) که پیش تر با آن آشنا شدیم، این مدل را آموزش داده و بهترین فرآپارامترها را گزارش کنید. (نیازی به آموزش مدل و انتخاب فرآپارامترها با آزمون و خطا نیست).

فاز پنجم: Support Vector Machine

SVM یا ماشین بردار پشتیبان یکی از روش های یادگیری ماشین برای طبقه بندی و رگرسیون است. این مدل با استفاده از یک خط یا صفحه ای که بین داده های دو دسته مختلف قرار می گیرد و فاصله بین این خط و نزدیک ترین نمونه ها از هر دو دسته را به حداکثر می رساند، کار می کند. این رویکرد به ایجاد یک مرز واضح بین دسته ها کمک می کند که امکان پیش بینی دقیق تر را فراهم می آورد. SVM برای داده هایی که دارای ویژگی های بسیاری هستند و نیازمند تفکیک دقیق تری می باشند، بسیار مفید است. در این تمرین قصد آن را داریم با مفهوم این روش بیشتر آشنا شویم.

43. به چه نقاطی support vector گفته می شود و آن را روی مثالی دلخواه نمایش دهید.

44. به نظر شما طبقه بند SVM برای طبقه بندی چه نوع داده هایی مناسب نیستند؟

45. درباره kernel ها و نقش آن ها در طبقه بندی توضیح دهید. (توضیح دهید وظیفه kernel ها چیست و چجوری به طبقه بندی کمک می کنند)

46. تفاوت soft svm classifier با hard svm classifier را بیان کنید.

47. نحوه استفاده از SVM در مسائل رگرسیون رو را با کشیدن شکل توضیح دهید.

48. حال مراحل زیر را برای پیاده سازی این روش انجام دهید :

1. با استفاده از کتابخانه های موجود با 2 کرنل RBF و Linear داده های خود را دسته بندی کنید.
2. ماتریس Confusion و هم چنین معیار های ارزیابی مدل مثل F1 , accuracy , Recall, ... را گزارش کنید و تحلیل خود را در گزارش ذکر کنید.
3. از کدام یک از روش های Grid search و Random search در اینجا بهتر است استفاده کنیم ؟
4. حال با استفاده از دو روش Random Search و Grid Search به ترتیب برای بازه دلخواه و مقادیر دلخواه و برای 2 کرنل RBF و Linear بهترین طبقه بند خود را پیدا کنید. (مجاز به استفاده از کتابخانه می باشید).

فاز ششم: ارزیابی مدل‌ها

معیارهای زیادی برای سنجش و ارزیابی عملکرد مدل‌ها وجود دارد. ارزیابی مدل‌های دسته‌بندی در یادگیری ماشینی به معنای ارزیابی عملکرد و کارایی مدل‌های مختلف است که برای دسته‌بندی داده‌ها استفاده می‌شوند. ارزیابی مدل‌های دسته‌بندی از اهمیت بسیاری برخوردار است زیرا به ما کمک می‌کند تا بتوانیم مدلی که می‌سازیم را با دقت بیشتری پیشرفت دهیم و اطمینان حاصل کنیم که عملکرد آن بهینه است. با استفاده از این معیارها و ارزیابی‌کننده‌های دیگر می‌توان مدل‌های دسته‌بندی را مقایسه کرده و انتخاب بهترین مدل را برای مسئله خاص خود انجام داد.

49. ماتریس درهم‌ریختگی⁶ چیست؟ از آن در چه مواردی استفاده می‌شود؟

50. ماتریس درهم‌ریختگی را برای تمامی مدل‌های classification پیاده‌سازی شده برای دادگان train و test به طور جداگانه رسم نموده، هر کدام از مدل‌ها را با مدل‌های دیگر مقایسه کنید.

51. در مورد underfitting و overfitting تحقیق کنید. سپس با توجه به نتایج بدست آمده از ماتریس درهم‌ریختگی دادگان train و test احتمال overfitting را برای هر یک از مدل‌ها بررسی کنید.

52. در این تمرین از شما می‌خواهیم برای ارزیابی مدل خود از ۴ معیار Accuracy, Precision, Recall و F1Score استفاده کنید. ابتدا درباره هر یک از ملاک‌های ارزیابی فوق تحقیق کنید، نحوه محاسبه هر یک را ذکر کرده و بیان کنید هر یک در چه شرایطی حائز اهمیت بیشتری هستند. مثال بزنید. برای مطالعه بیشتر می‌توانید از [این](#) لینک یا [این](#) لینک استفاده کنید.

53. تفاوت میان Precision و Recall را بیان کنید و توضیح دهید چرا هر کدام به تنهایی برای ارزیابی مدل کافی نیست؟ برای هر یک مثالی بیاورید که این مقادیر بالا هستند اما مدل عملکرد خوبی ندارد.

54. معیار F1 از چه نوع میانگین‌گیری استفاده می‌کند؟ تفاوت این نوع میانگین‌گیری با میانگین‌گیری عادی چیست و در اینجا چرا اهمیت دارد؟

55. همانطور که میدانید، در این مسئله بیش از ۲ کلاس داریم، در مورد Multi-class Metrics تحقیق کنید و سه حالت میانگین‌گیری Macro و Micro و Weighted را شرح دهید. و برای هر یک از این مدل‌ها هر 3 مقدار را محاسبه کنید. برای مطالعه می‌توانید از [این](#) لینک استفاده کنید.

56. درباره شاخص حساسیت (Sensitivity) و خصوصیت (Specificity) تحقیق کنید و بیان کنید که هر یک در چه مواقعی قابل استفاده هستند. (نیازی به پیاده‌سازی و محاسبه این موارد وجود ندارد).

بخش اختیاری: ROC Curve

ROC Curve یک نمودار ارزیابی برای مدل‌های دسته‌بندی است که به تعیین کیفیت و عملکرد مدل در تفکیک داده‌ها کمک می‌کند. ROC Curve مخفف Receiver Operating Characteristic Curve است.

این نمودار به طور خاص مقدار عملکرد مدل را بر اساس تعداد تشخیص داده‌های صحیح (True Positive Rate) نسبت به تعداد تشخیص داده‌های نادرست (False Positive Rate) نشان می‌دهد. در واقع، ROC Curve نمایش دهنده تغییرات در تشخیص صحیح و نادرست مدل در سطوح مختلف آستانه (Threshold) است که برای تصمیم‌گیری در دسته‌بندی استفاده می‌شود.

برای درک بهتر این نمودار ابتدا به سوالات زیر پاسخ کوتاه دهید.

57. منحنی ROC چیست و چگونه می‌توان آن را تفسیر کرد؟ در چه صورتی این نمودار بیانگر عملکرد بهتر مدل است؟

⁶ Confusion matrix

58. مساحت زیر منحنی ROC چه اهمیتی دارد؟ درباره شاخص AUC تحقیق کنید و نحوه تفسیر و محاسبه آن را شرح دهید.
59. چگونه می‌توان از منحنی ROC برای تعیین بهترین آستانه تصمیم‌گیری برای مدل دسته‌بندی استفاده کرد؟
60. روشی برای رسم منحنی ROC در حالت چند کلاسه ارائه دهید.
61. در این بخش با توجه به راه حلی که برای کشیدن نمودار ROC برای حالت چند کلاسه ارائه دادید منحنی ROC را برای هر قسمت رسم کنید و برای هر کلاس AUC را محاسبه کرده و نتیجه را گزارش کنید. همچنین تحلیل خود را بر روی عملکرد هر یک با توجه به معیار ROC در گزارش ذکر کنید.
62. در صورتی که دادگان به طور چشمگیری نامتوازن باشند، چه راهکارهایی برای مدیریت این موضوع پیشنهاد می‌کنید؟ برای رفع این مشکل چگونه از ROC استفاده می‌کنیم؟
63. اکثر مدل‌های دسته بندی قابلیت دریافت یک توزیع احتمالاتی برای کلاس‌ها را دارند. از هر یک از مدل‌ها به جای دریافت کلاس پیش‌بینی شده احتمال قرار گرفتن در کلاس‌ها را بدست آورده و سپس threshold یا حد آستانه قرارگیری در هر کلاس را تغییر دهید. نتایج را با نتایج قبل مقایسه کنید. آیا تغییر حدود آستانه به طور یکنواخت بر تعداد مشاهدات پیش بینی شده در هر دسته تاثیر گذار است؟

نکات پایانی

- توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. از ابزارهای تحلیل داده مانند نمودارها استفاده کنید.
- تمیزی پیاده‌سازی کدها، استفاده مناسب از شیء‌گرایی در پیاده‌سازی مدل‌ها و توابع و دسته بندی منظم آن‌ها حائز نمره امتیازی خواهد بود.
- پس از مطالعه کامل و دقیق صورت پروژه، در صورت وجود هرگونه ابهام یا سوال با طراحان پروژه در ارتباط باشید.
- نتایج، گزارش و کدهای خود را در قالب یک فایل فشرده با فرمت AI_CA3_[stdNumber].zip در سامانه ایلرن بارگذاری کنید.
- محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.
- دقت کنید که نیازی به آپلود مجموعه داده‌ها در سامانه ایلرن نیست.